RSS čtečka s inteligentní filtrací

Projekt z umělé inteligence

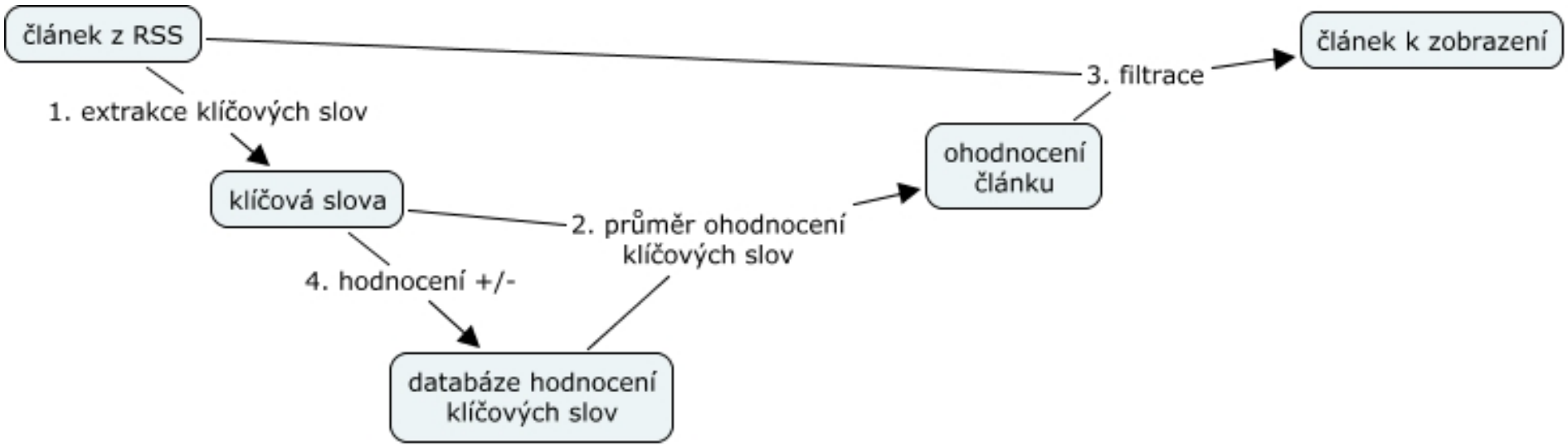
Jiří Procházka

FI MUNI

4. dubna 2013

Aktuální stav implementace

- Více-uživatelský systém pro perzistenci dat ✔
- Extrakce klíčových slov ✔
- Průměr ohodnocení klíčových slov ✔
- Hodnocení klíčových slov ✔
- Server interface ✔
- Klient interface ✗
- UI, filtrace ✗



Extrakce klíčových slov

Nakonec pomocí malé knihovny topia.termextract - rychlé, spolehlivé a účinné:

> *This package determines important terms within a given piece of content. It uses linguistic tools such as Parts-Of-Speech (POS) and some simple statistical analysis to determine the terms and their strength.*

Db

Spotlight - rozsáhlý a náročný software pro vlastní instalaci, použití jako služby (bez instalace) je riskantní.
Moje vlastní experimenty s NLTK nad korpusy s frekvenční distribucí kolokací a samostatných slov byly celkem úspěšné, ale topia.termextract extrahuje n-gramy i pro větší n, a nemá problémy s velmi krátkými texty, což RSS popisy bývají.

```
MINE:  ['Nokia', 'VP8', 'Google', 'HTC', 'Internet', 'licensing', 'infringement', 'declara-
tion']
TOPIA: ['8', 'Nokia', 'VP', 'patent']
TOPIAf: [')." Nokia', '8', 'IPR declaration', 'Internet Engineering Task Force list-
ing 64', 'June 14.', 'Lists Patents Google', 'MPEG LA', 'March 8', 'Nokia', 'Nokia re-
serves', 'Nokia-Google patent news', 'VP', 'VP 8', 'VP 8 format', 'VP 8-
related trial', 'VP 8.', 'import ban', 'license deal', 'patent', 'patent applica-
tions', 'patent infringement allegations', 'patent infringe-
ment cases', 'sales bans', 'video codec patent claims']
MINE:  ['robotic', 'Hitachi', 'car']
TOPIA: ['car']
TOPIAf: ['New submitter terrywo 5', 'Tiny Robo-
Taxi Carries 1 Passenger', 'car', 'drop passengers']
MINE:  ['ICANN', 'domain', 'names', 'list']
TOPIA: ['domain', 'domain names', 'name']
TOPIAf: ['27 domain names', 'Arabic names', 'Assigned Names', 'Internet Corpora-
tion', 'Mozaic domains', 'Regional Winners', 'Top Level Domain plans', 'domain', 'do-
main names', 'name', 'non-English domain names']
MINE:  ['Internet', 'proposal', 'Senate', 'bill', 'sales', 'final', 'tax']
TOPIA: ['Internet', 'sale', 'tax']
TOPIAf: ['2014 budget bill', 'Dick Durbin', 'Illinois Democrat', 'Internet', 'Inter-
net sales tax', 'Internet sellers', 'Senators Mike Enzi', 'Wyoming Republi-
can', 'sale', 'sales tax', 'sales tax bill', 'senate yesterday', 'tax']
MINE:  ['Asian', 'Easter', 'steady', 'shares']
TOPIA: ['Asian', 'Asian shares']
TOPIAf: ['Asian', 'Asian markets', 'Asian shares', 'Easter holidays', 'Hong Kong']
MINE:  ['Sweeping', 'welfare', 'changes']
TOPIA: []
TOPIAf: ['benefit cuts', 'welfare system']
```

Průměr ohodnocení klíčových slov

Používám depth limited search. Váha ohodnocení hypernym klesá se vzdáleností od klíčového slova v hierarchii podle funkce $f(x) = 1/x$:

Problém: přidané stejně hodnocené hypernymum *snižuje* ohodnocení celého textu $\rightarrow$ použití koeficientů na zmírnění tohoto účinku:
$depthRating(rating, depth, Q) = rating \cdot (depth + 1)^{-Q}$
Pro výpočet dělitele v průměru se použije tento vzorec s rating 1.0 a druhým hloubkovým koeficientem:

pole (slovo,rating,hloubka)	Q=1.0, normalní průměr	Q=1.0,Qd=1.0	Q=1.0,Qd=0.5
[('rudá',1.0,0)]	1.0	1.0	1.0
[('rudá',1.0,0),('barva',1.0,2)]	0.6667	1.0	0.8453

Problémy k řešení

Implementace klientské části, filtrace a GUI elementů v Tiny Tiny RSS.

Konec prezentace

Děkuji za pozornost